Read & Improve: A Novel Reading Tutoring System

Rebecca Watson
iLexIR Ltd
Cambridge
United Kingdom
bec@ilexir.co.uk

Ekaterina Kochmar
Dept of Computer Science
University of Bath
United Kingdom
ek762@bath.ac.uk

ABSTRACT

We introduce a new readability tutoring system, Read & Improve, a freely available online resource aimed at supporting learners of English and English Language Teaching (ELT) professionals by improving English learners' reading proficiency. Using a combination of machine learning approaches and natural language processing techniques, Read & Improve detects learning needs of every student and makes sure no learner is left behind by identifying reading content at an appropriate level of *readability* and helping learners acquire new words through accessible dictionary definitions and content exploration functionality.[1]

1. INTRODUCTION

Reading is one of the fundamental language skills. Developing this skill is an essential part of language acquisition, both for native speakers and second language learners [9, 13]. At the same time, developing reading ability takes a considerable amount of time, and, as any learning process, it gets interrupted if readers lose motivation [8, 15]. Such factors as not having a range of engaging reading content offered and being presented with reading material at the wrong level of readability are some of the major contributors to the decreased motivation in readers [11]. In addition to language learners themselves, English Language Teaching (ELT) professionals face similar problems, as finding engaging reading content at the right level of readability is a challenging and a time-consuming task. In this paper, we present Read and Improve (*R&I*), a freely available, open-access educational system that is aimed at both language learners and teachers.[2]

[1]This work has been done while the second author was a Senior Research Associate at the University of Cambridge. We thank Cambridge English for supporting this research via the ALTA Institute. We are also grateful to the anonymous reviewers for their valuable feedback.

[2]`https://readandimprove.englishlanguageitutoring.com/`

To ensure that the reading content provided to a learner is at an appropriate level of readability, *R&I* uses machine learning methods described in [18] to automatically label texts with readability levels corresponding to the Common European Framework of Reference for Languages (CEFR) [6]. The CEFR is an international standard that describes language ability on a six-point scale from A1 for beginners level up to C2 for advanced level of language proficiency.

To ensure that the reading content presented to a learner is engaging, *R&I* employs news articles that are sourced from news websites in real time. To source news content, *R&I* monitors both RSS Feeds from news websites and the publicly available Common Crawl News (`CC-NEWS`) Dataset.[3] A fully automated *Indexing Pipeline* (RIIP, herein) processes article URLs from the RSS Feeds and `CC-NEWS` files and automatically labels the *readability* of each article's text. News articles are generally available for learners on *R&I* within 10 minutes of publishing on an RSS news feed and in 3-6 hours of the article's publishing time if sourced from `CC-NEWS`. As compared to other domains, news articles have the additional benefit of being generally free of grammatical and spelling errors, which allows us to achieve more reliable linguistic analysis and to provide learners with high quality reading content. *R&I*'s user interface (UI) enables learners to not only read the latest news articles but also to perform keyword search to find articles on topics that they are interested in at their desired CEFR level(s).

A number of applications for various groups of readers, including native and non-native speakers, readers with cognitive impairments, and children, to name just a few, have been developed in recent years. In contrast to the previous work [13, 16, 17], our platform is aimed specifically at developing reading ability in non-native speakers of English. Our approach bears similarities to the Read-X [14] and REAP [10] systems, while also being actively developed and supported as an open-access educational platform available online. *R&I* is markedly different from other available applications, as in addition to providing text search functionality (as in [5]) and vocabulary acquisition help (as in [4]), it supports comprehension testing and personalisation.

The rest of this paper is structured as follows: Section 2 provides an overview of the system's architecture, Section 3 describes the current UI functionality, and finally Section 4

[3]`http://commoncrawl.org/2016/10/news-dataset-available/`

concludes the paper and describes future work.

2. SYSTEM ARCHITECTURE

Figure 1 illustrates the system architecture of *R&I*. We do not describe the full details of system components or cloud resources and configuration here, as this is outside the scope of the paper. Instead, we provide a general overview of the components and their use of natural language processing (NLP).

2.1 API

The API connects to an information retrieval index ('IR Engine'), a database ('DB Engine'), and several APIs to provide the data and search functionality required by the UI. The IR Engine employs Elasticsearch[4] (ES) and includes several distinct indices that facilitate search over news articles and other data.

2.2 RIIP

RIIP is responsible for processing articles into the ES article index. In order to prevent duplicate processing, the pipeline modules first check whether the output file(s) already exist in the 'Data Lake', a single store of all data processed. The API monitors the set of URLs listed in RSS feed(s) and the set of `CC-NEWS` files for new items, and if found, these are sent to RIIP for processing. Therefore, ingestion of new articles through the system requires no manual effort, and up-to-date news content is continuously processed and made available to learners via the UI.

RIIP modules include: the *Extractor*, that extracts text and other information from news articles (i.e. HTML); *RASP*, that parses the text to provide linguistic information [2];[5] the *LevelMarker* module, that labels the text for readability (on the CEFR scale); and finally the *ES* module that indexes text and other linguistic information.

2.3 LevelMarker Module

For RIIP's LevelMarker module we follow Briscoe et al. [3], and define the task of learning readability levels as a discriminative preference ranking task. We employ their machine learning (ML) software and use linguistic features outlined by Xia et al. [18] that represent a text's readability.

2.3.1 Data

We have crawled three publicly available news websites to create datasets: Breaking News English (BNE)[6] (2771 articles), News in Levels (NIL)[7] (6373 articles) and Tween Tribune (TT)[8] (7768 articles). These websites have news articles labelled in terms of their readability however each website's readability levels are based on different scales as shown in Table 1.[9] Each of these datasets are considered to be *parallel* as they contain multiple versions of the same articles simplified across different levels. While the BNE and NIL datasets are designed for L2 English learners, the TT is designed to help L1 learners (early and school-aged readers).

Table 1: Dataset levels and distributions.

(a) BNE

BNE level	CEFR level	Count
0	A2	386
1	A2	386
2	A2	386
3	A2-B1	418
4	B1-B2	392
5	B2	392
6	C1-C2	412

(b) NIL

NIL level	Count
1	2126
2	2124
3	2123

(c) CER

Exam	**CEFR level**	**Count**
KET	A2	64
PET	B1	60
FCE	B2	71
CAE	C1	67
CPE	C2	69

(d) TT

TT level	Count
Grade K-4 (0)	1965
Grade 5-6 (1)	2029
Grade 7-8 (2)	1771
Grade 9-12 (3)	2003

Table 2: 5-fold cross-validation tests for each dataset.

Source	**Pearson's**	**Spearman's**	**Kendall's**
BNE	0.8338	0.8368	0.6873
NIL	0.9217	0.9164	0.7880
TT	0.9055	0.9250	0.8071
CER	0.9155	0.9185	0.8015

2.3.2 Evaluation

RIIP employs a model trained on the full BNE dataset as this dataset can be reliably mapped to the CEFR scale (Table 1). Based on this mapping we determined the ranges of ML scores that corresponded to each CEFR level (using observed score range from training data). We tested our model on the Cambridge English Readability (CER) dataset,[10] a publicly available dataset of 331 texts spanning CEFR levels A2 to C2 [18]. On this test set, our model achieves 0.83 Pearson's, 0.85 Spearman's and 0.71 Kendal's correlation coefficient. We also ran 5-fold cross-validation for each dataset[11] and present the results in Table 2.

2.4 ES index

In addition to article index, we create 'WordInfo' and 'CALD' indexes. The CALD indexing system processes definitions from the Cambridge Advanced Learner's Dictionary (CALD) to populate the CALD index. The LexDoop system employs Hadoop[12] to process the Data Lake files (currently around 1 million articles) to produce raw frequency counts of linguistic properties for every word lemma.[13] Following this step, these lemma statistics are collated and added to the 'WordInfo' index.

[4] `https://www.elastic.co/products/elasticsearch`

[5] `https://ilexir.co.uk/rasp/index.html`

[6] `https://breakingnewsenglish.com/`

[7] `https://www.newsinlevels.com/`

[8] `https://www.tweentribune.com/`

[9] BNE to CEFR level map provided by the website: `https://breakingnewsenglish.com/news_levels.html`

[10] `https://ilexir.co.uk/datasets/index.html`

[11] We split the data randomly into training and test sets, ensuring an even distribution of class labels.

[12] Apache Hadoop: `https://hadoop.apache.org/`

[13] LexDoop is also used to process (any number of) `CC-NEWS` files in parallel, as required.

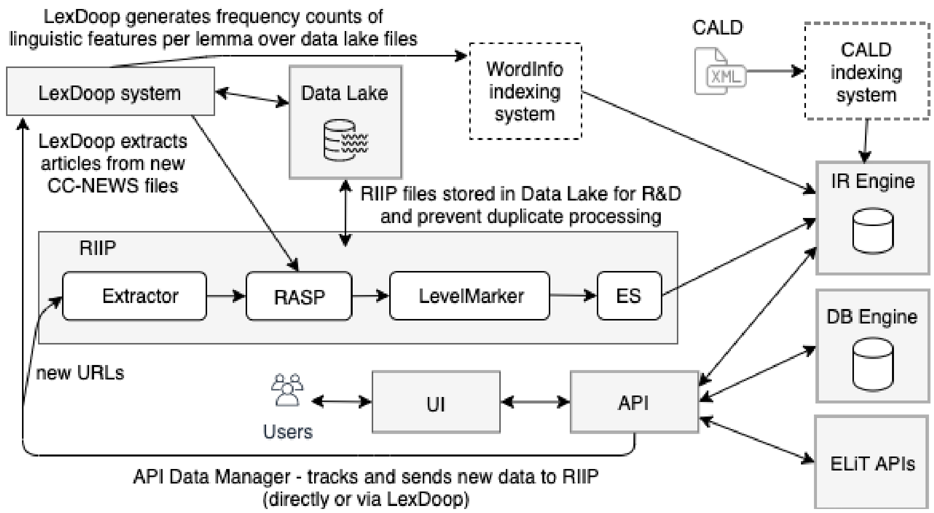


Figure 1: Overview of *R&I* architecture. *R&I* is hosted within, and relies upon, cloud computing services from Amazon Web Services (AWS). Components that use cloud AWS services are shown with grey backgrounds.

2.5 Sanitisation

To make sure the content provided on the platform is acceptable for a wide range of readers across various ages and cultures, we apply content *"sanitisation"* strategy, whereupon we automatically filter out news articles that contain words pertaining to the topics that might be considered offensive in some cultures or inappropriate for younger readers. The list of around 1600 such taboo words was curated using the lists of taboo words from social media. Sanitisation is run within RIIP and the API and, in case the sanitisation system makes an error, the UI enables admin users to mark articles as 'unsafe' (or vice versa).

3. READING ON THE PLATFORM

We define the *R&I* functionality in terms of four major aspects, which cover the tutoring system's ability to provide learners and teachers with engaging reading content at the appropriate level of readability (§3.1); help learners develop their vocabulary in English (§3.2); run comprehension tests (§3.3); and allow learners to revisit texts they read, words they clicked on and tests they submitted (§3.4).

3.1 Finding engaging reading material at an appropriate level

The first step for learners accessing *R&I* is to define their language proficiency level. Learners can log in to *R&I* using their account credentials from Write & Improve,[14] a freely available system linked to the reading platform, that is able to assess and provide feedback on a learner's writing proficiency. Once logged in, *R&I* defaults reading proficiency to current writing proficiency, but a learner can change their CEFR reading level.

Figure 2 contains a screenshot of the *search page*'s results showing the latest news articles at the learner's CEFR level (currently B1). The search page provides learners with snippet(s) of the article text, and they can click on any of the titles listed on this page in order to load the *article view page* where they can read the article itself. In addition, search by keywords is enabled on *R&I* to allow learners to find articles not only at their level of readability, but also on the topics of their interest.

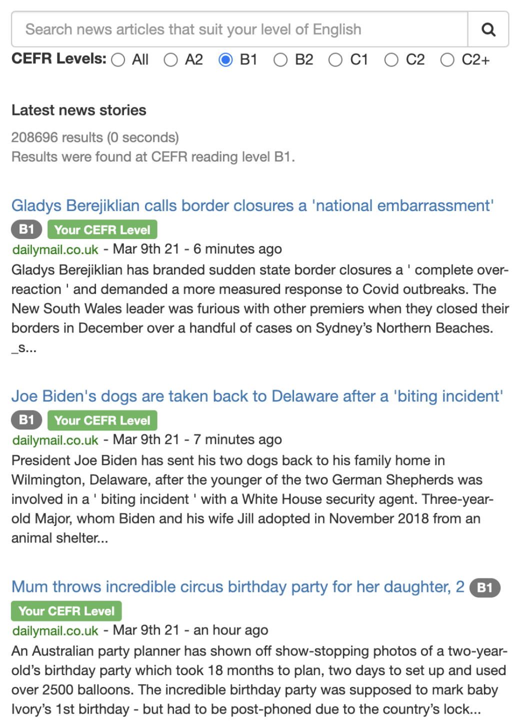
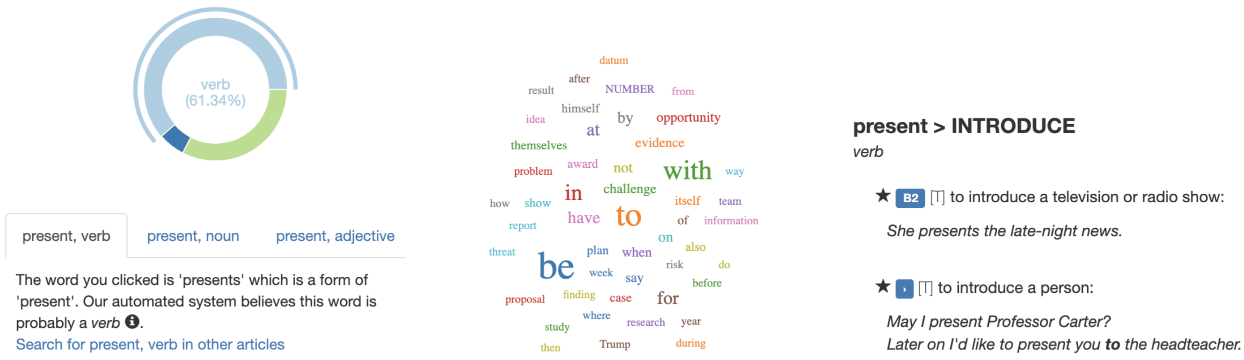


Figure 2: Screenshot: search results.

3.2 Developing one's vocabulary

Vocabulary is very important in language learning to the point that language learning itself would sometimes be equated

[14] `https://writeandimprove.com/`
R&I employs Write & Improve APIs developed by ELiT: `https://englishlanguageitutoring.com/`



(a) Part of speech (PoS) statistics (b) Word cloud (c) CALD word sense 'INTRODUCE'

Figure 3: Screenshots of the sections of the 'Word Information' and 'English Dictionary' panels on the UI. Here, the user clicked on the word *presents*, used as a verb. The pie chart in (a) illustrates the relative frequency of all PoS categories for the lemma *present* across all articles. The word clouds in (b) contain the 50 lemmas most frequently co-occurring with *present* as a verb in grammatical relations (where font size reflects relative frequency), and (c) shows the dictionary definition.

with knowing language vocabulary [12]. To help learners with vocabulary acquisition and development, *R&I* allows them to select any words they do not recognise or wish to learn more about within the article view page. When a learner clicks on an unknown word, *R&I*'s UI launches two side panels for *Word Information* and *English Dictionary* (shown in Figure 3) to display information available for the word in the 'WordInfo' and 'CALD' index, respectively, as described in §2.4.

Several searches can be performed by clicking on links within the *Word Information* panel and words within the co-occurrence word cloud. These links to search results shown in *R&I*'s search page enable learners to perform advanced, linguistically motivated searches intuitively and learn how vocabulary is used in context.

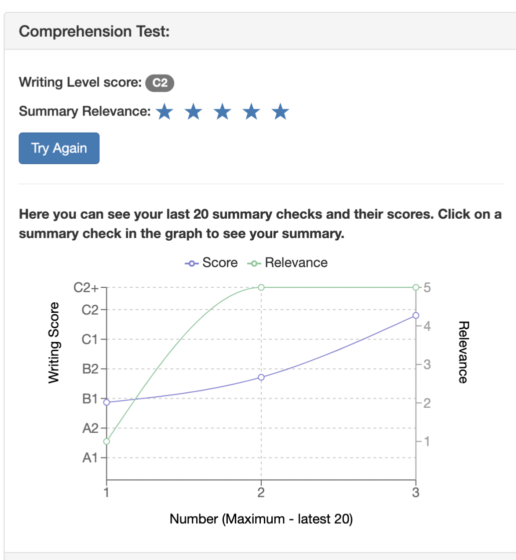


Figure 4: Screenshot: Comprehension Test panel. Learners are able to click on the graph to view previous summaries, which they can refine and re-submit.

3.3 Running comprehension tests

R&I allows users to submit a summary of the article as a *comprehension test* in the *Comprehension Test* panel on the article view page (Figure 4). *R&I* automatically scores these summaries and returns a *writing score*, determined by a mature feature-based automated essay scoring (AES) model [1, 3, 20], graded on the CEFR scale via the Write & Improve API, and a *relevance score* based on the maximum sentence-level cosine similarity value, which is then converted to a score in the range 0–5 using the lexical overlap between the article and the summary [7] that shows whether the learner captured the main salient topics in the article.

3.4 Accessing reading history

All history of learner interaction with the *R&I* platform, including texts, vocabulary items and submitted summaries is available to the learners on the personal *My Reading* pages.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented Read & Improve, a freely available, open-access reading tutoring system that is aimed at language learners and teachers. Currently, it is a prototype system, and thence most of its components will benefit from further research on the platform. For instance, we are planning to improve our Indexing Pipeline using quality human annotated training data and user analytics that we are collecting via the *R&I* platform.

R&I records learners' actions on the UI, which in turn, will provide valuable data for use in further research and development. For example, [19] employed the comprehension test data collected by the platform to develop a new automated comprehension test (summary assessment) marking system suitable for use in *R&I*. Further, each learner's data may be useful in directly improving their learning experiences. For example, analysis of an individual learner's history could be used to tailor custom content and testing for each learner. This symbiotic relationship, developed in an ecosystem of freely available educational system benefiting from cutting-edge research, will ultimately produce a state-of-the-art ELT resource.

5. REFERENCES

[1] Ø. E. Andersen, H. Yannakoudakis, F. Barker, and T. Parish. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[2] T. Briscoe, J. Carroll, and R. Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sydney, Australia, July 2006. Association for Computational Linguistics.

[3] T. Briscoe, B. Medlock, and Ø. Andersen. Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory, Nov. 2010.

[4] J.-J. Chen, C.-Y. Yang, P.-C. Ho, M. C. Tsai, C.-F. Ho, K.-W. Tuan, C.-T. Tsai, W.-B. Han, and J. S. Chang. Learning to Link Grammar and Encyclopedic Information to Assist ESL Learners. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 213–218. Association for Computational Linguistics, 2019.

[5] M. Chinkina, M. Kannan, and D. Meurers. Online Information Retrieval for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, pages 7–12. Association for Computational Linguistics, 2016.

[6] Council of Europe. Common European Framework of Reference for Languages: Learning, Teaching, Assessment, 2011.

[7] R. Cummins, H. Yannakoudakis, and T. Briscoe. Unsupervised Modeling of Topical Relevance in L2 Learner Text. In *BEA@NAACL-HLT*, 2016.

[8] Z. Dörnyei. Motivation in second and foreign language learning. *Language teaching*, 31(3):117–135, 1998.

[9] W. H. DuBay. The principles of readability. *Online Submission*, 2004.

[10] M. Heilman, L. Zhao, J. Pino, and M. Eskenazi. Retrieval of Reading Materials for Vocabulary and Reading Practice. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88. Association for Computational Linguistics, 2008.

[11] D. Hirsh and P. Nation. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, 8(2):689–689, 1992.

[12] C. James. *Errors in language learning and use: Exploring error analysis*. Routledge, 2013.

[13] N. Madnani, B. B. Klebanov, A. Loukina, B. Gyawali, P. L. Lange, J. Sabatini, and M. Flor. My turn to read: An interleaved e-book reading tool for developing and struggling readers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146, 2019.

[14] E. Miltsakaki and A. Troutt. Read-X: Automatic Evaluation of Reading Difficulty of Web Text. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 7280–7286. Association for the Advancement of Computing in Education (AACE), 2007.

[15] N. Oroujlou and M. Vahedi. Motivation, attitude, and language learning. *Procedia-Social and Behavioral Sciences*, 29:994–1000, 2011.

[16] L. Rello, R. Baeza-Yates, S. Horacio, S. Bott, R. Carlini, C. Bayarri, A. Gòrriz, S. Gupta, G. Kanvinde, and V. Topac. Dyswebxia 2.0!: Accessible text for people with dyslexia (demo). In *Proceedings W4A 2013, The Paciello Group Web Accessibility Challenge*, Rio de Janeiro, Brazil, 2013.

[17] Z. Weiss, S. Dittrich, and D. Meurers. A linguistically-informed search engine to identify reading material for functional illiteracy classes. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*, pages 79–90. Linköping Electronic Conference Proceedings 152, 2018.

[18] M. Xia, E. Kochmar, and T. Briscoe. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2016)*, pages 12–22, San Diego, California, June 2016. Association for Computational Linguistics.

[19] M. Xia, E. Kochmar, and T. Briscoe. Automatic learner summary assessment for reading comprehension. In *Proceedings of NAACL-HLT 2019*, pages 2532–2542, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] H. Yannakoudakis, T. Briscoe, and B. Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.